

Optimal trade-offs for pattern matching with k mismatches

Paweł Gawrychowski¹ and Przemysław Uznański²

¹Haifa University, Israel

²ETH Zürich, Switzerland

Abstract

Given a pattern of length m and a text of length n , the goal in k -mismatch pattern matching is to compute, for every m -substring of the text, the exact Hamming distance to the pattern or report that it exceeds k . This can be solved in either $\tilde{O}(n\sqrt{k})$ time as shown by Amir et al. [J. Algorithms 2004] or $\tilde{O}((m+k^2) \cdot n/m)$ time due to a result of Clifford et al. [SODA 2016]. We provide a smooth time trade-off between these two bounds by designing an algorithm working in time $\tilde{O}((m+k\sqrt{m}) \cdot n/m)$. We complement this with a matching conditional lower bound, showing that a significantly faster *combinatorial* algorithm is not possible, unless the combinatorial matrix multiplication conjecture fails.

1 Introduction

The basic question in algorithms on strings is pattern matching, which asks for reporting (or detecting) occurrences of the given pattern in the text. This fundamental question comes in multiple shapes and colors, starting from the exact version considered already in the 70s [6]. Here we are particularly interested in the approximate version, where the goal is to detect fragments of the text that are *similar* to the text. Two commonly considered variants of this question is pattern matching with k errors and pattern matching with k mismatches. In the former, we are looking for a fragment with edit distance at most k to the pattern, while in the latter we are interested in a fragment that differs from the pattern on up to k positions (and has the same length). The classical solution by Landau and Vishkin [7] solves pattern matching with k mismatches in $\mathcal{O}(nk)$ time for a text of length n . For larger values of k , Abrahamson [1] showed how to compute the number of mismatches between every fragments of the text and the pattern of length m in total $\mathcal{O}(n\sqrt{m \log m})$ time with convolution. Later, Amir et al. [2] combined both approaches to achieve $\mathcal{O}(n\sqrt{k \log k})$ time.

An obvious and intriguing question is what are the best possible time bounds for pattern matching with k mismatches. An unpublished result attributed to Indyk [3] is that, if we are interested in counting mismatches for every position in the text, then this is at least as difficult as multiplying boolean matrices. In particular, it implies that one should not hope to significantly improve on the $\mathcal{O}(n\sqrt{m})$ time complexity of an *combinatorial* algorithm. However, this is not sensitive to the bound k on the number of mismatches. In a recent breakthrough, Clifford et al. [4] introduced a new repertoire of tools and showed an $\mathcal{O}((k^2 \log k + m \text{polylog } m) \cdot n/m)$ time algorithm. In particular, this is near linear-time for $k = \mathcal{O}(\sqrt{m})$ and improves on the previous algorithm of Amir et al. [2] that runs in $\mathcal{O}(n/m \cdot (k^3 \log k + m))$ time.

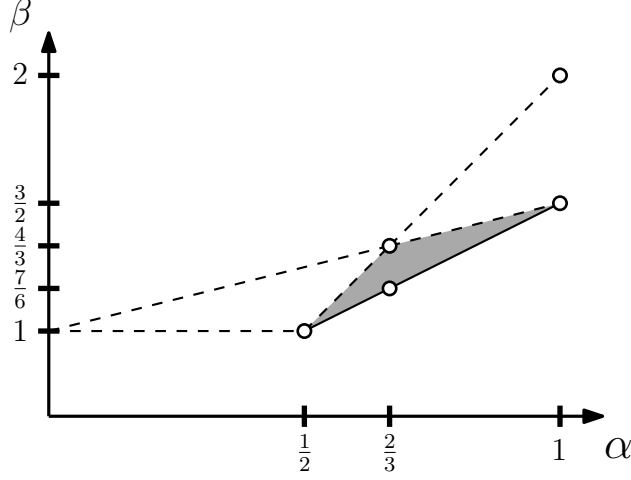


Figure 1: Running time $T = m^\beta$ on instances with $n = \Theta(m)$ and $k = m^\alpha$. Previous algorithms are represented by dashed lines and our algorithm is represented by solid line. For example, for $k = \Theta(m^{2/3})$ we improve the complexity from $\tilde{O}(m^{4/3})$ to $\tilde{O}(m^{7/6})$.

Results. We provide a smooth transition between the $\tilde{O}(n\sqrt{k})$ time algorithm of Amir et al. [2] and the $\tilde{O}((m + k^2) \cdot n/m)$ solution given by Clifford et al. [4]. The running time of our algorithm is $\tilde{O}((m + k\sqrt{m}) \cdot n/m)$. This matches the previous solution at the extreme points $k = \mathcal{O}(\sqrt{m})$ and $k = \Omega(m)$, but provides a better trade-off in-between. Furthermore, we prove that such transition is essentially the best possible. More precisely, we complement the algorithm with a matching conditional lower bound, showing that a significantly faster *combinatorial* algorithm is not possible, unless the popular combinatorial matrix multiplication conjecture fails.

Related work. Landau and Vishkin [7] solve pattern matching with k mismatches by checking every possible alignment with $k + 1$ constant-time longest common extension queries (also known as “kangaroo jumps”). The main idea in all the subsequent improvements is to use convolution, which essentially counts matches generated by a particular letter with a single FFT in time close to linear. Both Abrahamson [1] and Amir et al. [2] use convolution for letters often occurring in the pattern. Convolution is also used (together with random projections $\Sigma \rightarrow \{0, 1\}$ that can be derandomized with an extra $\mathcal{O}(\log n)$ factor) by Karloff [5] for approximate mismatches counting.

At a very high level, Clifford et al. [4] obtain the improved time complexity by partitioning both the pattern and the text into $\mathcal{O}(k)$ subpatterns and subtexts, such that the total number of blocks in their RLE is small. Resulting $\mathcal{O}(k^2)$ instances of RLE pattern matching with mismatches are then solved in $\mathcal{O}(k^2)$ total time, leading to an $\tilde{O}((k^2 + m) \cdot n/m)$ time algorithm for the original problem.

Overview of the techniques. We observe that the reduction from [4] can be done so that, instead of many small instances, we end up with a *single* new instance of $\mathcal{O}(k)$ -mismatch pattern matching. The resulting new pattern and text have RLE consisting of $\mathcal{O}(k)$ blocks and the problem is reduced to RLE pattern matching with k mismatches. Since for RLE pattern matching with mismatches there is a matching quadratic conditional lower bound (by reducing from the 3SUM problem), it might seem that no improvement here is possible without making a significant breakthrough.

We show that this is not necessarily the case, by leveraging that the RLE strings are compressed version of strings of $\mathcal{O}(m)$ length. Thus, letters that appear in only a few blocks of the compressed pattern can be treated in a fashion similar to [2] by producing a representation of all matches generated by a block in the compressed pattern against a block in the compressed text, in constant time per a pair of blocks. For letters that appear in many blocks, we can essentially “uncompress” the corresponding fragment of the pattern, and run the classical convolution, taking advantage of the fact that uncompressed versions are of length $\mathcal{O}(m)$. Setting threshold appropriately, we solve the obtained RLE pattern matching in time $\tilde{\mathcal{O}}(k\sqrt{m})$ time. All in all, we obtain an $\tilde{\mathcal{O}}((m + k\sqrt{m}) \cdot n/m)$ time solution to the original problem.

2 Upper bound

The goal of this section is to prove the following theorem:

Theorem 2.1. *k -mismatch pattern matching can be solved in time $\mathcal{O}(n/m \cdot (m \log^2 m \log |\Sigma| + k\sqrt{m \log m}))$.*

We begin with the standard trick of reducing the problem to $\lceil n/m \rceil$ instances of matching a pattern P of length m to a text T of length $2m$ and work with such formulation from now on. Therefore, the goal now is to achieve $\mathcal{O}(m \log^2 m \log |\Sigma| + k\sqrt{m \log m})$ complexity.

We start by highlighting the kernelization technique of Clifford et al. [4]. An integer $\pi > 0$ is an x -period of a string $S[1, m]$ if $\text{Ham}(S[\pi, m-1], S[0, m-1-\pi]) \leq x$ (cf. Definition 1 in [4]). Note that compared to the original formulation, we drop the condition that π is minimal from the definition.

Lemma 2.2 (Fact 3.1 in [4]). *If the minimal $2x$ -period of the pattern is ℓ , then the starting positions of any two occurrences with k mismatches of the pattern are at distance at least ℓ .*

The first step of algorithm is to determine the minimal $\mathcal{O}(k)$ -period of the pattern. More specifically, we run the $(1 + \varepsilon)$ -approximate algorithm of Karloff [5] with $\varepsilon = 1$ matching the pattern P against itself. This takes $\mathcal{O}(m \log^2 m \log |\Sigma|)$ time and, by looking at the approximate outputs for offsets not larger than k , allows us to distinguish between two cases:

- every $2k$ -period of the pattern is at least k , or
- there is a $4k$ -period $\ell \leq k$ of the pattern.

Then we run the appropriate algorithm as described below.

No small $2k$ -period. We again run Karloff’s algorithm with $\varepsilon = 1$, but now we match the pattern with the text. We look for positions i where the approximate algorithm reports at most k mismatches, meaning that $\text{Ham}(P, T[i .. i + m - 1]) \leq 2k$. By Lemma 2.2, there are $\mathcal{O}(m/k)$ such positions, and we can safely discard all other positions. Then, we test every such position using the “kangaroo jumps” technique of Landau and Vishkin [7], using $\mathcal{O}(k)$ constant-time operations per position, in total $\mathcal{O}(m)$ time.

$T' = \text{hokuspokusopensezame}$										$P = \text{abracadabra}$									
h	s	u	e	z		s	u	e	z	#	a	c	b	\$	\$				
o	p	s	n	a		p	s	n	a	#	b	a	r	\$	\$				
k	o	o	s	m		o	o	s	m	#	r	d	a	\$	\$				
u	k	p	e	e		k	p	e	e	#	a	a	\$	\$	\$				
$T^* = \text{hsuez opsna koosm ukpee suez\# psna\# oosm\# kpee\#}$										$P^* = \text{acb\$\$ bar\$\$ rda\$\$ aa\$\$\$}$									

Figure 2: Example of rearranging of text and pattern, with parameter $\ell = 4$.

Small $4k$ -period. Let $\ell \leq k$ be any $4k$ -period of the pattern. For a string S and $0 \leq i < \ell$, let $\{S\}_{\ell,i} = S[i]S[i+\ell]S[i+2\ell] \dots$ up until end of S . We denote by $\{S\}_\ell$ an ℓ -encoding of S , that is the string $\{S\}_{\ell,1}\{S\}_{\ell,2} \dots \{S\}_{\ell,\ell-1}$. Let $\text{runs}(S)$ be the number of runs in S . Denote $\text{runs}_\ell(S) = \sum_{i=1}^{\ell} \text{runs}(\{S\}_{\ell,i})$, and observe that it upperbounds the number of runs in $\{S\}_\ell$.

Lemma 2.3 (Lemma 6.1 in [4]). *If P has a $4k$ -period not exceeding k , then $\text{runs}_\ell(P) \leq 5k$.*

We proceed with the kernelization argument. Let T_L be the longest suffix of $T[0, m-1]$ such that $\text{runs}_\ell(T_L) \leq 6k$. Similarly, let T_R be the longest prefix of $T[m, 2m-1]$ such that $\text{runs}_\ell(T_R) \leq 6k$. Let $T' = T_L T_R$. Obviously, $\text{runs}_\ell(T') \leq 12k$.

Lemma 2.4 (Lemma 6.2 in [4]). *Every $T[i, i+m-1]$ that is an occurrence of P with k mismatches is fully contained in T' .*

Thus we see that k -mismatch pattern matching is reduced to a kernel where the ℓ -encoding of both the text and the pattern have few runs, that is, compress well with RLE.

From now on assume that both T' and P are of lengths divisible by ℓ . If it is not the case, we can pad them separately with at most $\ell - 1 < k$ characters each, not changing the complexity of our solution. Let m_1 and m_2 be integers such that $m_1 \cdot \ell = |T'|$ and $m_2 \cdot \ell = |P|$, $m_1 \geq m_2$.

We rearrange both P and T' to take advantage of their regular structure. That is, we define $T^* = \{T'\}_\ell \{T''\}_\ell$, where $T'' = T'[\ell+1, m_1 \cdot \ell] \#^\ell$. Observe that T^* is a word of length $2m_1 \cdot \ell$, composed first of m_1 blocks of the form $T'[i]T'[i+\ell] \dots T'[i+(m_1-1)\ell]$ for $0 \leq i < \ell$, and then of m_1 blocks of the form $T'[i+\ell] \dots T'[i+(m_1-1)\ell] \#$.

Similarly, we define $P^* = \{P \$^{(m_1-m_2)\ell}\}_\ell$. Again we observe that P^* is the word of length $m_1 \cdot \ell$, composed of blocks of the form $P[i]P[i+\ell] \dots P[i+(m_2-1)\ell] \$^{m_1-m_2}$ for $0 \leq i < \ell$. Example of this reduction is presented on Figure 2.

Next we show that T^* and P^* maintain the Hamming distance between any possible alignment of T' and P .

Lemma 2.5. *For any integer $0 \leq \alpha \leq (m_1-m_2)\ell$, let $x = \lfloor \alpha/\ell \rfloor$ and $y = \alpha \bmod \ell$. Let $\beta = x+y \cdot m_1$. Then*

$$\text{Ham}(T'[\alpha, \alpha + m_2 \cdot \ell - 1], P) = \text{Ham}(T^*[\beta, \beta + m_1 \cdot \ell - 1], P^*) - (m_1 - m_2) \cdot \ell.$$

Proof. Observe that

$$\text{Ham}(T'[\alpha, \alpha + m_2 \cdot \ell - 1], P) = \sum_{i=0}^{m_2-1} \sum_{j=0}^{\ell-1} \delta(T'[x\ell + y + i\ell + j], P[i\ell + j]), \quad (1)$$

where δ is indicator of character inequality. Observe that $P[i\ell + j] = P^*[i + j \cdot m_1]$, for $0 \leq j < \ell - y$ there is $T'[x\ell + y + i\ell + j] = T^*[(x+i) + (y+j)m_1]$, and for $\ell - y \leq j < \ell$ there is $T'[x\ell + y + i\ell + j] = T''[(x+i)\ell + (y+j-\ell)] = T^*[(x+i) + (y+j-\ell)m_1 + \ell m_1] = T^*[(x+i) + (y+j)m_1]$. Additionally, for $m_2 \leq i < m_1$, $P^*[i + j \cdot m_1] = \$$, which always generates a mismatch with any character in T^* . Thus

$$\begin{aligned} (1) &= \sum_{i=0}^{m_2-1} \sum_{j=0}^{\ell-1} \delta(T^*[(x+i) + (y+j)m_1], P^*[i + j \cdot m_1]) = \\ &= -(m_1 - m_2)\ell + \sum_{i=0}^{m_1-1} \sum_{j=0}^{\ell-1} \delta(T^*[(x+i) + (y+j)m_1], P^*[i + j \cdot m_1]), \quad \square \end{aligned}$$

We see that it is enough to find all occurrences of P^* in T^* with $(k + (m_1 - m_2) \cdot \ell)$ mismatches, where $k + (m_1 - m_2)\ell \leq 2k$, $|P^*| = |T'| \leq m$ and $|T^*| = 2|T'| \leq 2m$. Additionally, $\text{runs}(P^*) \leq 5k + \ell \leq 6k$ and $\text{runs}(T^*) \leq 12k + \ell \leq 13k$.

Now we describe how to solve the kernelized problem exactly (where we count matches/mismatches for all possible alignments, not just detect occurrences with up to k mismatches), using the stated properties of T^* and P^* .

Consider a letter $c \in \Sigma$. For a string S , we denote by $\text{runs}(S, c)$ the number of runs in S consisting of occurrences of c . Fix a parameter t . Call a letter c such that $\text{runs}(P^*, c) > t$ a heavy letter, and otherwise call it light. Now we describe how to count the number of mismatches for each type of letters. This is reminiscent to a trick originally used by Abrahamson [1] and later refined by Amir et al. [2].

Heavy letters. For every heavy letter c separately we use a convolution scheme. Since both P^* and T^* are of size $\mathcal{O}(m)$, this takes time $\mathcal{O}(m \log m)$ per every such letter. Since $\sum_{c \in \Sigma} \text{runs}(P^*, c) = \text{runs}(P^*) \leq 6k$, there are $\mathcal{O}(k/t)$ heavy letters, making the total time $\mathcal{O}(mk \log m/t)$.

Light letters. First, we preprocess P^* , and for every light letter c we compute a list of runs consisting of occurrences of c . Our goal is to compute the array $A[0, |T^*| - |P^*|]$, where $A[i]$ counts the number of matching occurrences of light letters in $T^*[i, i + |P^*| - 1]$ and P^* .

We scan T^* , and for every run of a particular light letter, we iterate through the precomputed list of runs of this letter in P^* . Observe that, given a run of the same letter in P^* and in T^* , matches generated between $T^*[u, v]$ and $P^*[y, z]$ account for a piecewise linear function. More precisely, for all integer $u \leq i \leq v$ and $y \leq j \leq z$, we need to increase $A[i - j]$ by one. To see that we can process pair of runs in constant time, we work with discrete derivatives, instead of original arrays.

Given sequence F , we define its discrete derivate DF as follow: $(DF)[i] = F[i] - F[i - 1]$. Correspondingly, if we consider generating function $F(x) = \sum_i F[i]x^i$, then $(DF)(x) = F(x) \cdot (1 - x)$ (for convenience, we assume that arrays are indexed from $-\infty$ to ∞).

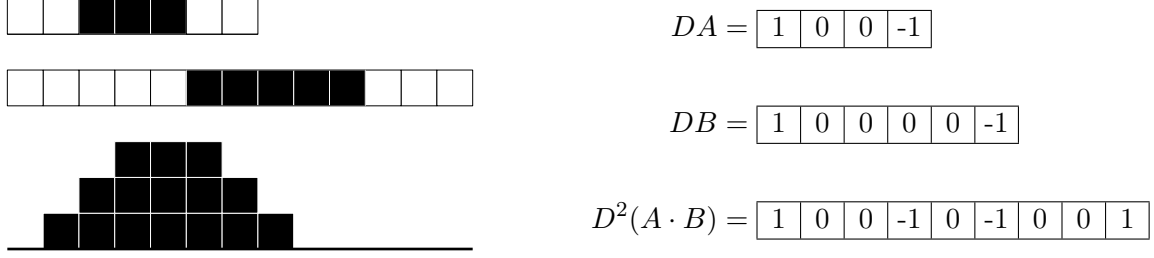


Figure 3: On the left - a run in the pattern and a run in the text (both represented by black boxes) consisting of the same character and a histogram of the matches they generate. On the right - first derivatives of the indicator arrays and second derivate of the match array, without padding zeroes.

Now consider indicator sequences $T_{u,v}[i] = \mathbf{1}(u \leq i \leq v)$ and $P_{y,z}[j] = \mathbf{1}(-z \leq j \leq -y)$. To perform the update, we set $A[i+j] += T_{u,v}[i] \cdot P_{y,z}[j]$ for all i, j , or simpler using generating functions:

$$A(x) += T_{u,v}(x) \cdot P_{y,z}(x), \quad (2)$$

where $T_{u,v}(x) = \sum_{i=u}^v x^i$ and $P_{y,z}(x) = \sum_{j=y}^z x^{-j}$. However, we observe that $DT_{u,v}$ and $DP_{y,z}$ have particularly simple forms: $DT_{u,v}(x) = x^u - x^{v+1}$ and $DP_{y,z}(x) = x^{-z} - x^{-y+1}$. Thus it is easier to maintain second derivate of A , and (2) becomes:

$$D^2 A(x) += x^{u-z} - x^{v-z+1} - x^{u-y+1} + x^{v-y+2}.$$

All in all, we can maintain $D^2 A$ in constant time per pair of runs, or in $\mathcal{O}(k \cdot t)$ total time, since every list of runs is of length at most t , and there are at most $13k$ runs in T^* . Additionally, in $\mathcal{O}(m)$ time we can compute $A[0]$ and $A[1]$, allowing us to recover all other $A[i]$ s from the formula $A[i] = (D^2 A)[i] + 2A[i-1] - A[i-2]$.

Setting $t = \sqrt{m \log m}$ gives the total running time $\mathcal{O}(k\sqrt{m \log m})$ in both cases as claimed.

3 Lower bound

Below we present a conditional lower bound, which expands upon an idea attributed to Indyk [3]. Main idea here is to use rectangular matrices instead of square, and use the padding accordingly. However, we pad using the same character in both text and pattern, increasing the number of mismatches only by a factor of 2.

Recall the combinatorial matrix multiplication conjecture stating that, for any $\varepsilon > 0$, there is no *combinatorial* algorithm for multiplying two $n \times n$ boolean matrices working in time $\mathcal{O}(n^{3-\varepsilon})$. The following formulation is equivalent to this conjecture:

Conjecture 3.1 (Combinatorial matrix multiplication). *For any $\alpha, \beta, \gamma, \varepsilon > 0$, there is no combinatorial algorithm for multiplying an $n^\alpha \times n^\beta$ matrix with an $n^\beta \times n^\gamma$ matrix in time $\mathcal{O}(n^{\alpha+\beta+\gamma-\varepsilon})$.*

The equivalence can be seen by simply cutting the matrices into square block (in one direction) or in rectangular blocks (in the other direction).

Now, consider two boolean matrices, A of dimension $M' \times N$ and B of dimension $N \times M$, for $M' \geq M \geq N$. We encode A as text T , by encoding elements row by row and adding some padding. Namely:

$$T = \#^{M^2} r_1 \#^{M-N+1} r_2 \#^{M-N+1} \dots \#^{M-N+1} r_{M'} \#^{M^2}$$

where $r_i = r_{i,1} \dots r_{i,N}$ and $r_{i,j} = 0$ when $A_{i,j} = 0$ and $r_{i,j} = j$ when $A_{i,j} = 1$. Similarly, we encode B as P column by column, using padding shorter by one character:

$$P = c_1 \#^{M-N} c_2 \#^{M-N} \dots \#^{M-N} c_M$$

where $c_j = c_{1,j} \dots c_{N,j}$ and $c_{i,j} = 0'$ when $B_{i,j} = 0$ and $c_{i,j} = i$ when $B_{i,j} = 1$.

Observe that, since we encode 0s from A and B using different symbols, and encoding of 1s is position-dependent, r_i and c_j will generate a match only if they are perfectly aligned and there is k such that $r_{i,k} = c_{k,j}$, or equivalently $A_{i,k} = B_{k,j} = 1$. Since each block (encoded row plus following padding) is either of length $N + 1$ for rows or N for columns, there will be at most one pair row-column aligned for each pattern-text alignment.

The total number of mismatches, for each alignment, is at most $2NM$ (since there are at most MN non-# text characters that are aligned with pattern, and at most MN non-# pattern characters). We can recover whether any given entry of $A \cdot B$ is a 1, since if so the number of mismatches for the corresponding alignment is decreased by 1.

We have $|T| = \Theta(M'M)$ and $|P| = \Theta(M^2)$. By setting $M = \sqrt{m}$, $M' = \frac{n}{\sqrt{m}}$ and $N = \frac{k}{\sqrt{m}}$ we have the following:

Corollary 3.2. *For any positive $\varepsilon, \alpha, \kappa$, such that $\frac{1}{2}\alpha \leq \kappa \leq \alpha \leq 1$ there is no combinatorial algorithm solving pattern matching with $k = \Theta(n^\kappa)$ mismatches in time $\mathcal{O}((k\sqrt{m} \cdot n/m)^{1-\varepsilon})$ for a text of length n and a pattern of length $m = \Theta(n^\alpha)$, unless Conjecture 3.1 fails.*

If we denote by $\omega(\alpha, \beta, \gamma)$ the exponent of fastest algorithm to multiply a matrix of dimension $n^\alpha \times n^\beta$ with a matrix of dimension $n^\beta \times n^\gamma$, we have:

Corollary 3.3. *For any positive $\varepsilon, \alpha, \kappa$, such that $\frac{1}{2}\alpha \leq \kappa \leq \alpha \leq 1$ there is no algorithm solving pattern matching with $\Theta(n^\kappa)$ mismatches in time $\mathcal{O}(n^{\omega(2-\alpha, 2\kappa-\alpha, \alpha)/2-\varepsilon})$ for a text of length n and a pattern of length $\Theta(n^\alpha)$.*

References

- [1] Karl R. Abrahamson. Generalized string matching. *SIAM J. Comput.*, 16(6):1039–1051, 1987.
- [2] Amihood Amir, Moshe Lewenstein, and Ely Porat. Faster algorithms for string matching with k mismatches. *J. Algorithms*, 50(2):257–275, 2004.
- [3] Raphaël Clifford. Matrix multiplication and pattern matching under Hamming norm. <http://www.cs.bris.ac.uk/Research/Algorithms/events/BAD09/BAD09/Talks/BAD09-Hammingnotes.pdf>. Retrieved March 2017.
- [4] Raphaël Clifford, Allyx Fontaine, Ely Porat, Benjamin Sach, and Tatiana A. Starikovskaya. The k -mismatch problem revisited. In *SODA*, pages 2039–2052. SIAM, 2016.
- [5] Howard J. Karloff. Fast algorithms for approximately counting mismatches. *Inf. Process. Lett.*, 48(2):53–60, 1993.
- [6] Donald E. Knuth, Jr. James H. Morris, and Vaughan R. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977.
- [7] Gad M. Landau and Uzi Vishkin. Efficient string matching with k mismatches. *Theor. Comput. Sci.*, 43:239–249, 1986.